8-16-2016

High-Level Modeling for Computer-Aided Clinical Trials of Medical Devices

Houssam Abbas
University of Pennsylvania, habbas@seas.upenn.edu

Zhihao Jiang
University of Pennsylvania, zhihaoj@seas.upenn.edu

Kuk Jin Jang
University of Pennsylvania, jangkj@seas.upenn.edu

Marco Beccani
University of Pennsylvania, beccani@seas.upenn.edu

Jackson Liang
Hospital of University of Pennsylvania, jackson.liang@uphs.upenn.edu

See next page for additional authors

High-Level Modeling for Computer-Aided Clinical Trials of Medical Devices

Author(s)
Houssam Abbas, Zhihao Jiang, Kuk Jin Jang, Marco Beccani, Jackson Liang, Sanjay Dixit, and Rahul Mangharam

High-Level Modeling for Computer-Aided Clinical Trials of Medical Devices

Houssam Abbas* and Zhihao Jiang* and Kuk Jin Jang*
and Marco Beccani* and Jackson Liang† and Rahul Mangharam*
*Department of Electrical and Systems Engineering
University of Pennsylvania Philadelphia, PA 19104
Email: habbas, zhihaoj, jangkj, beccani, rahulm@seas.upenn.edu
†Cardiovascular Division
Hospital of the University of Pennsylvania
jackson.liang@uphs.upenn.edu

***Abstract*—Medical devices like the Implantable Cardioverter Defibrillator (ICD) are life-critical systems. Malfunctions of the device can cause serious injury or death of the patient. In addition to rigorous testing and verification during the development process, new medical devices often go through *clinical trials* to evaluate their safety and performance on sample populations. Clinical trials are costly and prone to failure if not planned and executed properly. Evaluating devices on computer models of the relevant physiological systems can provide helpful insights into the safety and efficacy of the device, thus helping to plan and execute a clinical trial. In this paper, we demonstrate how to develop high-level physiological models of cardiac electrophysiology and how to apply them to the Rhythm ID Head to Head Trial (RIGHT), a 5-year long clinical trial for comparing two ICDs. We refer to this as a Computer-Aided Clinical Trial (CACT). We explored two modeling options, a white-box model capturing the mechanisms of the physiological behaviors, and a black-box model which uses machine learning methods to synthesize physiological input signals. Both models were able to generate physiological inputs to the ICDs and we discuss the challenges and appropriateness of the two modeling options.**

I. Introduction

Medical devices have been developed to save and improve people's lives. In the U.S. for example, 10,000 people receive an Implantable Cardioverter Defibrillator (ICD), a heart rhythm adjustment device, every month [1]. The diagnostic and therapeutic functions of medical devices are becoming more autonomous in order to deliver timely therapy and reduce human labor and error. This increasingly sophisticated functionality is fulfilled by increasingly complex software. Both the software and hardware of a medical device are prone to errors and faults, and the failure modes depend as much on the design as on the patient's lifestyle and unique physiology. Malfunctions of the medical device can cause serious injury or death of the patient, so these life-critical devices are subject to much greater regulatory scrutiny and liability than consumer electronics hardware and software.

The *clinical trial* is a major difference between the development processes of consumer electronics and high-risk medical devices. In a typical trial, a group of patients that are treated with the new device are compared to a group of patients who are treated with the current standard of care (e.g., a different device currently on the market). The objective is to see whether the different devices result in significantly different effects on the patients. Consumer electronics might undergo some amount of field testing (e.g. a phone company may hand out prototypes of its latest phone to employees and monitor them), but it is insignificant compared to a clinical trial[1].

A. Device Verification

Verification activities take up most of the time in a typical hardware (HW) and software (SW) development process (up to 70% for HW by some estimates). Each verification activity has a sign-off criterion in the Test Plan, indicating that the design can proceed to the next phase of the development cycle. Such criteria will usually include successful linting and sufficiently high coverage metrics (code, functional, data and assertion coverage in particular). Model checking of certain sub-systems is also performed, where the sub-systems are chosen based on their criticality and their size. Finally, integration testing is performed when the chip is assembled together.

These activities are aimed at verifying that the design obeys its specification, and the input sequences fed to it during testing are aimed at that goal. In particular, there is no codified attempt at replicating the statistics or characteristics of 'real-world' scenarios (at least, not for functionality verification). In fact, the notion of 'real-world' scenarios already assumes a certain level of abstraction, typically that of a virtual prototype that can execute the software stack, and most verification activities listed above happen at the RT level or somewhat higher.

For medical devices, the real-world scenario is provided by the real world: the new ICD, say, is implanted in the trial population, and its efficacy is evaluated at the end of the trial, while its safety is evaluated throughout the trial. While a medical device's HW and SW can undergo the above verification activities, their functionality is not considered verified *until they have passed the clinical trial*. That's because, (in addition to the safety considerations), the human physiology

[1]Here we ignore post-market data collection since a trial happens *before* the new device can be brought to market

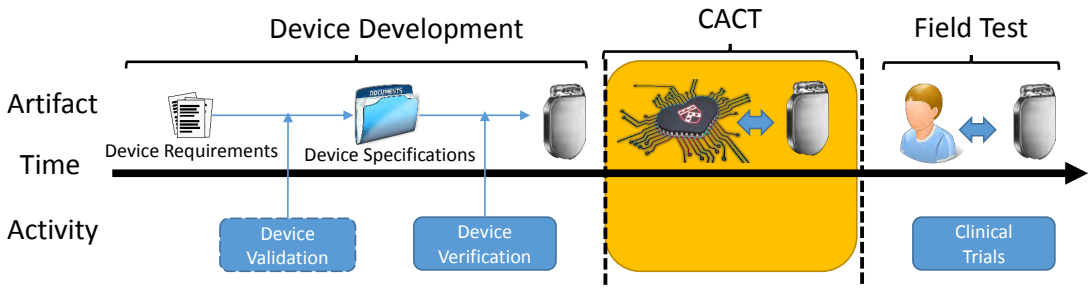


Fig. 1. The separation between medical device development and field test phrase

is so varied that debugging-oriented testing cannot possibly be comprehensive enough, or give statistically meaningful estimates of device efficacy.

In this paper we present the argument for Computer-Aided Clinical Trials (CACTs): a process in which high-level models of the human physiology are used to evaluate device performance on *virtual* trial populations, and the results of which are used to help plan a real clinical trial. In Section II we explain why randomized testing and data sets are insufficient in a CACT, and why physiology models are needed. We also discuss the appropriate abstraction level for models in a CACT. In Section III we give an overview of RIGHT, a 5-year long trial that compared two ICDs, and for which we conducted a Computer-Aided Clinical Trial retrospectively to validate our approach. Sections IV and V present the white-box and black-box models we built for use in the CACT, and in Section VI we discuss the implications of using such models for the verification activities above.

II. The need for input models

Why are models of the device's input needed for CACTs? Some life-critical devices like ICDs store the input signals they measure during an arrhythmia episode. Device companies likely have access to this (anonymized) physiological data. Isn't it then possible and sufficient to replay these input signals to the Design Under Test (DUT), thus foregoing the need for a model?

The need for physiological input models can be attributed to three reasons. We will illustrate these reasons using our running example of an ICD whose input is a 3D real-valued time series, known as the *electrogram*, or EGM.

Taming input complexity. The space of physiological input signals is complicated and with no evident structure. Formally, the input space is uncountably infinite since the physiological signal is real-valued, unlike the input to, say, a network router which contains discrete values. Moreover, the structure of the input space is far from obvious, unlike a network packet which has a well-defined, human-engineered structure. In fact, for a medical device, it is not even clear what is a valid input and what is not: given a signal, and given the immense variability of physiologies, how to tell automatically whether it could've been produced by a human heart or not? Modeling is essential to get a good handle on these aspects and tame the complexity of the input space. By modeling, we obtain a finite representation of the input space, impose a structure on it, and obtain a test of what is a physiologically valid signal (one that can be produced by the model) and what is not.

Separation of design and validation. Medical device companies likely have access to a vast set of data that is retrieved from their devices. This data might then be used to develop and test the device. Because a CACT is meant to be an independent assessment of the devices performance, the data used to develop it cannot be re-used in a CACT. A CACT that re-uses the development data will likely show very good performance and bias our estimate of true performance. Moreover, the data stored by the device might not be in a format that can be re-played on the DUT. For example, an ICD only stores arrhythmia episodes. On their own, these do not constitute a complete input signal to test the ICD.

Paucity of data. Physiological data is not readily available. By physiological data we mean the signals that are measured by a medical device and which it uses to diagnose the state of the patient and apply appropriate therapy. What data is available is usually siloed in proprietary platforms. Even regional medical centers don't have ready access to that data. E.g., the arrhythmia episodes that are recorded by an ICD can be viewed and printed, but not necessarily downloaded in digital form. Moreover, this data must be manually examined and labeled by a physician before it can be used to test a device. This provides a very strong motivation for the development of simulation models that can generate labeled input signals to the device.

A. Level of abstraction of the input

It is well-established that the higher the level of the abstraction, the easier it is to design system inputs and the easier it is to interpret test results. In HW validation, high-level modeling (and associated high-level testing) refers variably to design descriptions in SystemC or SystemVerilog, architectural descriptions in C/C++, or higher level virtual prototypes that can execute software meant to run on the HW under test.

The highest possible level of input to a MD is the patient's physiological state. However, the latter is not rigorously defined in the way that the state of an SoC is. When evaluating the power performance of an SoC, the SoC state is either ON, DROWSY, IDLE, OFF, etc, and each register is initialized to

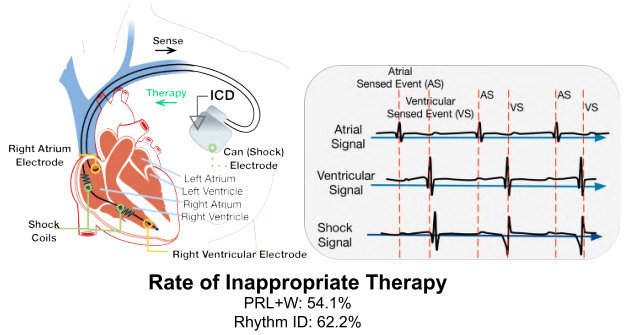


Fig. 2. ICD. Reported rates of inappropriate therapy are on a per-episode basis, as reported in RIGHT

a known value. What is the relevant state of a patient? For an ICD, the state obviously includes the presence of scars in the heart's muscle, history of arrhythmia, age and gender. Should we include the specific scar pattern? What about economic status, a failing kidney and a history of smoking? Moreover, each of these state variables is itself hard to capture. Obtaining a map of scar tissue in the heart requires an MRI scan, and the history of arrhythmia is qualitative at best. To turn this information into an operational input to the device, a lot of work has yet to happen both on the level of fundamental science and the level of data mining science.

The second abstraction level, one lower than patient state, is the description of physiological signals measured by the device. For an ICD, these are the electrograms. Inputs at a lower level than that don't make sense for medical device SW validation, and are extremely hard to interpret at best. This is the level at which we build our models of device inputs.

III. Case Study: Clinical Evaluation of VT/SVT Discrimination in ICDs

Before describing the models we built, we give a quick overview of ICD functionality and RIGHT. An ICD is an implantable device designed to treat ventricular arrhythmia, a life-threatening condition due to irregular electrical activity of the heart. A dual chamber ICD has two leads inserted into the heart against the wall of right atrium and right ventricle, respectively (Fig. 2). The leads measure local electrical activations of the heart which are referred to as electrogram (EGM) signals. Based on the EGM signals, the ICD diagnoses the heart condition and delivers therapy during ventricular tachycardia in the form of electrical pacing or shock to restore the normal heart rhythm.

Due to limited observability, the ICD algorithm may not have enough information to discriminate between life-threatening ventricular tachycardia (VT) and non-fatal supra-ventricular tachycardia (SVT). Diagnosing an SVT as VT leads to inappropriate shocks. Inappropriate therapy increases patient stress and is linked to increased morbidity. Depending on the particular ICD and its settings, the rates of inappropriate therapy can reach 62% of all delivered therapy episodes [2]!

A. RhythmID Goes Head-to-head Trial (RIGHT)

ICD manufacturers have developed algorithms to distinguish VT from SVT in order to reduce the chance for inappropriate therapy during SVT. The key constraint is that the ICD must always deliver therapy during a potentially fatal VT. To evaluate the performance of the algorithms, various clinical trials have been conducted. One particular trial was conducted to compare RhythmID, a VT/SVT discrimination algorithm developed by Boston Scientific, and PRLogic, an algorithm developed by Medtronic, in terms of time to first inappropriate therapy. The trial is named RhythmID Goes Head-to-head Trial (RIGHT). The initial assumption of the trial is that Rhythm ID is 25% better. The trial enrolled approximately 2000 patients and lasted 5 years from 2006 to 2011. However, at the end of the trial, it turned out that PRLogic is 27% better [2], resulting in trial failure.

Assume we are in 2006 during the planning of the trial. Can we use computer models of the heart's electrophysiology to conduct a Computer-Aided Clinical Trial (CACT) and provide useful insight that can prevent the trial failure, or lead to a re-planning of the trial?

B. Requirements for Physiological Models for Computer-Aided RIGHT

Physiological models generate input signals to a device. Therefore, they should be specifically developed for the device they will be connected to, and for the CACT they are meant to be used in. This poses a different set of restrictions on the input model than that imposed by the debugging activities listed above.

- For RIGHT, the objective is to determine whether an algorithm can make the right VT vs SVT decision: the model will feed signals to the device, but the device will not feed signals back to the model. Thus an *open-loop model* is appropriate.
- The algorithm should be evaluated on a virtual population that shares certain key characteristics with real populations. Thus there is no need for full input space coverage, since 'corners' of the input space might be rare conditions.[2]
- The signals generated by the model need to be physiologically plausible, so we can make statistical inferences based on them. Evaluating device performance on a set of inputs half of which could not possibly be generated by a human heart is meaningless. Therefore, the generated signal set *must conform to the statistics of available real data sets*.
- Finally, at the CACT stage, model interpretability is not as important as during design verification, since the goal is no longer to re-play bugs and figure out their cause.

Fig. 3 illustrates the importance of these different aspects in CA RIGHT, compared with device V&V.

[2]There may be interest in observing the rare conditions, but the point is that this isn't strictly needed *for the CACT*.

	Device V&V	CA RIGHT
Closed-loop	★★★★★	★★
Input Coverage	★★★★★	★★★
Model Validation	★★	★★★★
Model Identification	★★★	★★★★
Model Interpretability	★★★★★	★★

Fig. 3. Importance for different aspects

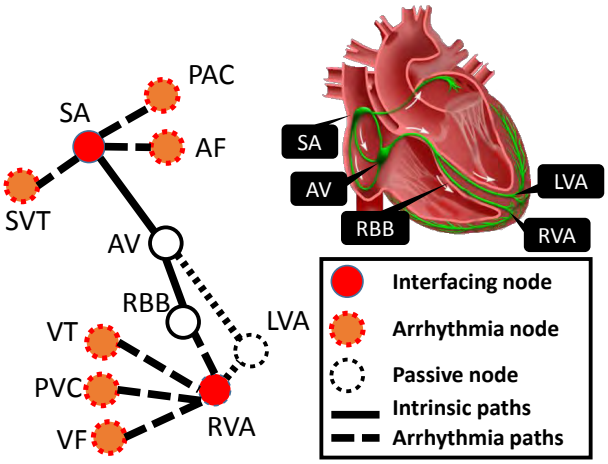


Fig. 4. Model of electrical conduction system of the heart

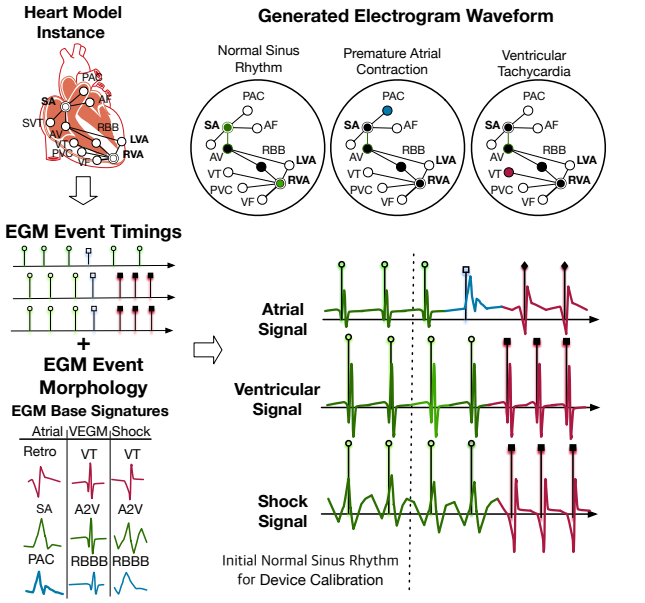


Fig. 5. EGM generation. The timing model produces the events boolean signal (top left) and the base EGM signatures are overlaid on the events (bottom left) to produce the final signal (bottom right). Different arrhythmias are modeled by different timing models (top right).

Given that a high-level model is essential for generating the devices input signals, we now ask: what kind of model should we develop?

IV. White Box Models of Electrophysiology

A white-box model captures the *mechanisms* for generating input behaviors. In the context of ICD operation, electrical depolarizations initiate from various locations of the heart, and conduct through the electrical conduction system of the heart. When the heart tissue where ICD leads are implanted is depolarized, the electrical voltage changes are monitored by the ICD leads. The resulting waveforms are referred to as electrogram, or EGM, signals, and constitute the input signals to the ICD. A white box model for EGM signals should capture the mechanisms that influence features of the EGM signals, which in turn affect device behaviors.

One of the challenges for developing white-box models is the level of abstraction used for modeling. What is the minimum amount of detail required to distinguish features of the EGM signals that affect device behaviors? In case of a dual chamber ICD, the features are the timing and morphology of the EGM signals, which are determined by the generation and conduction of electrical depolarizations within the heart. We developed a white-box heart model for EGM generation, which is based on the principles of clinical electrophysiology. As shown in Fig. 4, each solid circle represents a specific anatomical location of the heart. In this case, we only model the anatomical locations that can affect the electrical behaviors between the atrial lead and the ventricular lead of the ICD. For each location, a state machine models the timing behaviors of the generation and blocking of electrical depolarizations. The solid lines in Fig. 4 represent the electrical connections among anatomical locations of the heart. For each solid line, a state machine models the timing delay between two locations. We also model different source of abnormal electrical depolarizations and their connections with the main model structure, which are represented as dotted circles and lines in Fig. 4.

This model structure generates the timing of electrical depolarizations for both atrial and ventricular channels. According to the source of the electrical depolarizations, an EGM morphology is overlayed on top of the timing events, which completes EGM generation. The EGM morphologies are collected from EGM signals of real patients. The EGM generation process is demonstrated in Fig. 5.

For each heart condition, we obtained physiological ranges for the timing parameters from the clinical literature [3], and built a synthetic cohort by uniformly sampling from the ranges. We were able to construct 600 synthetic heart models for each of 19 heart conditions, and simulated the heart models to obtain 11,000+ arrhythmia episodes. The EGMs are then fed into our implementations of RhythmID and PRLogic algorithms to evaluate their rates of inappropriate therapy.

A. CACT Result With White-box Models

We implemented the two discrimination algorithms based on the available literature.

1) Rate of Inappropriate Therapy Across Various Populations: The obtained rates of inappropriate detection were

6.65% for Rhythm ID and 2.91% for PRL+W ($P < 0.0001$) on a per-episode basis, assuming an equal number of patients from each arrhythmia in the synthetic cohort. The corresponding relative improvement *of PRL+W over Rhythm ID* is 56%. In other words, the in-silico trial reveals that PRL+W algorithm differentiates between VT and SVT better than RhythmID. Our findings are consistent with the observations of the RIGHT trial itself [2]. We also varied the distribution of the arrhythmias in the synthetic cohort, and re-computed the cohort-wide rates of inappropriate therapy. As an example, Fig. 6 shows the results for the uniform distribution and a distribution that approximates that of RIGHT's cohort [2, Table 1]. It can be seen that indeed, PRL+W maintains a better rate of arrhythmia discrimination across the board.

2) Effect of Device Parameter Setting: In HW verification, different versions of the same HW are tested. E.g., a single core, dual core, and quad core versions of a processor may be tested. Or, different memory sizes' impact on latency and power consumption may be evaluated. Each version is targeted towards a different market segment.

Analogously, ICDs have a number of parameters which can be tuned by the physicians to accommodate specific patient conditions. One of the main causes of VT/SVT misclassifications is inappropriate parameter setting [4]. For the physicians to set appropriate parameters, it is very important to understand how the change of one parameter can affect the discriminating capability of the algorithm. With CACT, one can subject the *same* synthetic population to different settings of the parameters at virtually no cost. This is impossible to do with a real trial, since a patient cannot be implanted with two devices. In the CA RIGHT we evaluated the effect of two ICD parameters on specificity and sensitivity of PRL+W. The first parameter is the ***duration*** of arrhythmia before the ICD makes a therapy decision. The parameter for PRL+W is the number of consecutive fast ventricular intervals which can be set from 8 to 20 beats. In this experiment we explore the values {8,10,12,16,18,24,30} . From the results (Fig. 7) we observe that the specificity increases monotonically with the length of the duration, which matches the intuition as the device can examine a longer history of the arrhythmia episode

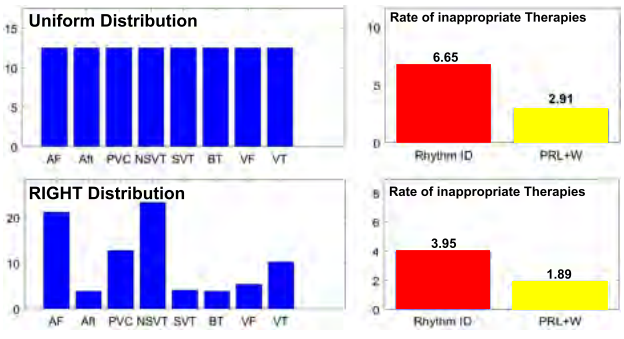


Fig. 6. Rate of inappropriate detection (2^{nd} column) for different arrhythmia distributions (1^{st} column). The upper-left distribution is uniform, and the lower-left distribution is that of the baseline characterization in RIGHT [2]. The x-axis lists the simulated arrhythmias.

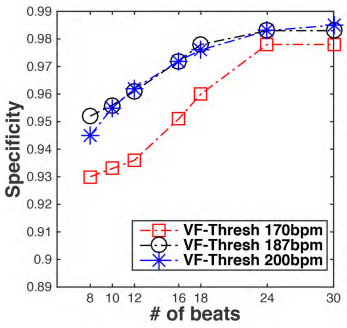


Fig. 7. Effect of Duration and VF threshold params on Specificity

with longer duration, and also allows a greater chance for the arrhythmia to self-terminate. This can prevent inappropriate detections therefore prevent inappropriate therapies.

The second parameter we varied is the ***VF threshold***. If the ventricular rate is faster than the VF threshold for a period of time the ICD will confirm detection and deliver therapy without going into the SVT/VT discrimination algorithm. In this experiment we explored the values {170,184,200} msec. As the value increases, specificity also increases as more episodes are examined by the SVT/VT discrimination algorithm.

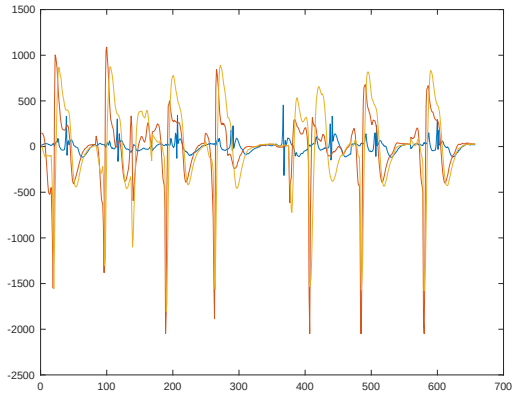
The result of of changing the two parameters matches the clinical results in [5] which further confirms the usefulness of CACT.

B. Discussion

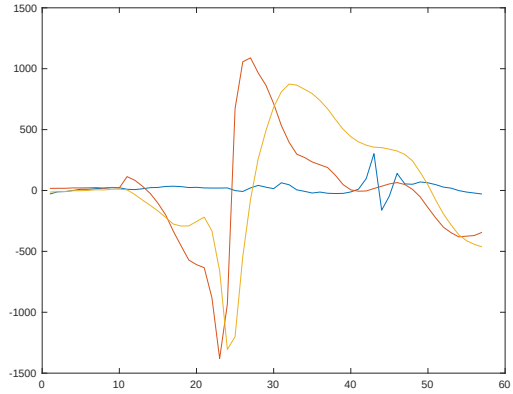
The advantage of white-box modeling in the current context is two-fold: first, by construction, it provides the sequence of events, in the heart, that led to the generation of a particular signal. In our example, the model shows how electrical signals conduct throughout the heart, which can distinguish VT from SVT. Second, by changing parameters that have a physiological meaning, white-box models enable the simulation of clinically relevant rare scenarios.

However, white-box models have a major disadvantage, which is the inability to identify joint distributions of the model's parameters from patient data. Our EGM-generating white box model has over 30 timing parameters. Ranges for *individual* timing parameters can be found in the clinical literature [3]. However, the joint distribution of these parameters is not available. Identifying them during a clinical procedure is theoretically possible, but practically impossible. Not being able to identify the joint distribution of the parameters diminishes model validity and its capability to represent specific patient groups. Both of these aspects are important for CACT. Moreover, white-box modeling also requires a significant amount of domain-specific and device-specific knowledge to construct.

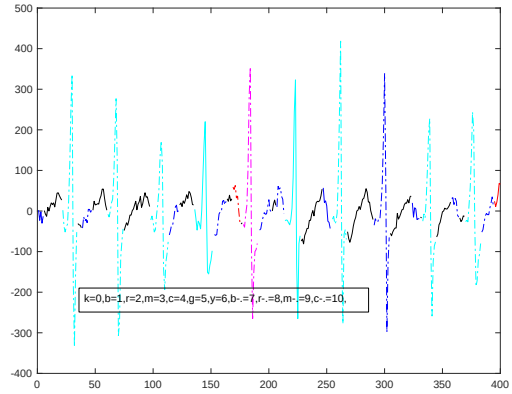
These disadvantages do not prevent the use of white-box modeling during device evaluation. For example, the diabetes model developed in University of Virginia is a white-box model. FDA accept simulation results of the model as a substitute for animal trials, which saves significant time and money, and reduces animal testing.



(a) Electrogram (one color per channel)



(b) Sample motif (one color per channel)



(c) A (one-dimensional) time series, which is used for learning an HMM model. The legend indicates 11 motifs were learned ($k = 0, \ldots, 10$). Different segments of the time series are colored with the motif that best explains them. Note how repetitions of motif 10, for example (Cyan-dash-dot) are not identical, but display variations.

Fig. 8. Electrogram for learning black box model, sample motif, and one channel of the electrogram overlaid with the learned motifs.

V. Black box models of electrophysiology

In contrast to a white box model, a black box model does not seek to model the mechanisms that generate the physiological signals. Rather, it directly models the emergent phenomenon that is the physiological signals. In the context of ICD verification, for example, the input to the ICD consists of a 3-dimensional time series $y_t \in \Re^3$, $t \in \{t_1, \ldots, t_m\}$, called the electrogram. See Fig. 8. Following [6], such a time series can be modeled as being generated by a Hidden Markov Model (HMM). An HMM is a finite state machine whose transitions are probabilistic: the probability of transitioning between states x_i and x_j is given by A_{ij}, $1 \leq i, j \leq N$. Matrix A is the *transition probability matrix*. In our case, the time series is modeled as being generated by the repetition of a finite number of *motifs* s^k, $k = 1, \ldots, K$. Each motif is a short duration signal $s^k : [0, T_k] \rightarrow \Re^3$. See Fig. 8 (b). The repetitions of a motif are not identical: two parameters, ϕ and ρ, enable time-varying modifications, as follows. A state of the HMM is given by $x = (k, \rho, \phi)$. At each time step t, if the current state of the HMM is $x_t = (k_t, \rho_t, \phi_t)$, then the model produces the value $y_t = \phi_t \cdot s^{k_t}(\rho_t) + \nu_t$, where s^k is the k^{th} motif, ϕ_t is the scaling factor, $\rho_t \in [0, T_k]$ is the time of a motif sample, and ν_t is a stochastic noise value. By sampling different values of k along the time series, different motifs are used (and repeated). With each repetition, the values ϕ and ρ serve to slightly alter the appearance of the motif.

The motifs s^k and transition matrix A are *learned from data* using Expectation-Maximization (EM), a classical soft clustering algorithm. The data used for learning is a set of electrograms, each of which is labeled by the arrhythmia and patient that produced it. This data is much more readily available than the data needed to fit timing parameters to the white-box model. Briefly, EM has two steps: in the Expectation (E) step, a matrix A and set of motifs s^k are given, and we infer the most likely sequence of HMM states $\mathbf{x} = (x_t)_{t \leq N}$ that explains the time series $\mathbf{y} = (y_t)_{t \leq N} : \mathbf{x} = argmax_{\mathbf{x}} P(\mathbf{y} \mid \mathbf{x})$. In the Maximization (M) step, $\mathbf{x}$ is used to update the motifs and the matrix A. The E and M steps are iterated until convergence. EM is usually implemented using Viterbi decoding for the E step, and various descent algorithms for the M step. See [7] for an excellent introduction to HMMs and EM. Fig.

Thus, for a given patient, and for a given arrhythmia, we can learn a black-box *generative* model. This model is then used to produce more exemplars of that arrhythmia, while obeying key statistical characteristics of real patient data. To generate a synthetic electrogram, we simply transition through the learned HMM using the transition probability matrix A. At time t, the state x_t is used to produce y_t in the manner described above. Just like the white box model, the generated electrograms are used as input to the device under test. A black box model is attractive because it can be learned from available data, thus automatically providing a measure of validity (i.e. confidence that when we simulate it, it will produce physiologically valid signals). It can also be simulated in real-time or faster, thus allowing for fast testing. And under suitable restrictions, if an appropriate device model is available, it could even be model checked.

One disadvantage of a black box model is its weak interpretability. The HMM bears no relation to physiological mechanisms such as ionic channels, muscle contractions, or wave propagation. The parameters of the model dont neces-

	White-box	Black-box
Closed-loop	★★★★	★
Input Coverage	★★★★	★★
Model Validation	★★	★★★
Model Identification	★★★	★★★★
Model Interpretability	★★★★	★★

Fig. 9. Ability to perform tasks with either type of model. More stars indicates it is easier to perform that task. Thus it is easier to do closed-loop testing with a white box model than it is with a black box model.

sarily have any physiological meaning. E.g. the conduction delay parameters of the automaton model of Section IV can be traced to the physiology of the heart generating the signal. In the HMM model on the other hand, the parameters include the state transition probability matrix A, which does not have a clear physiological interpretation. Thus if the CACT reveals poor device performance, we usually cant correlate that to any underlying physiological state of the patient, since this aspect is not part of the model (except to the degree that we know something about the population on whose data the model was trained to begin with). The doctors, on the other hand, are most interested in the patients physiological state (see discussion in Section II-A), and use the electrograms as a proxy for that. The lack of physiological interpretation also prevents the black-box model to be used in any closed-loop CACTs in which the physiological model is required to respond to device output.

VI. Discussion

In the last two sections, we discussed the two modeling options that we explored for a Computer-Aided RIGHT (CA RIGHT). Fig. 9 compares the ability of both models to capture the aspects discussed in Fig. 3. From the comparison we can see that the black-box model is more suitable for CA RIGHT while the white-box model is more suitable to perform closed-loop V & V. It is essential to determine the most suitable model during the planning of a CACT to minimize the effort of model revisions.

Generally, developing physiological models for a CACT includes answering the following questions:

1) Is the CACT a closed-loop trial or an open-loop trial?
2) If the CACT is open-loop, what are the characteristics of the input signals?
3) If the CACT is closed-loop, what are the physiological mechanisms that affect the input signals and respond to device outputs?
4) What is the minimum model that can capture the characteristics and/or the physiological mechanisms?
5) How valid is the model?
6) Is there patient data available with enough quality and quantity to identify the parameters of the physiological model?

Once a model has been developed and used to run a CACT for the DUT, can't it also be used to test the next generation of the device? Indeed it can (though this implies that it cannot be used in any future CACTs - see **Separation of design and validation** in Section II above). Using physiological models as we described them does pose some unique challenges when measuring medical device performance, and we briefly outline some of these in this section.

SW (and HW) sign-off is usually dependent on several coverage metrics exceeding pre-set levels. For example, code coverage needs to be above, say, 90%. Some of these metrics, like code coverage, are relatively removed from the abstraction level of the high-level physiological inputs (e.g, electrograms). If code coverage is too low, it may not be obvious what new input signals would increase it. Inputs to a medical device are more similar to a video stream used to test a video processing application than to a test sequence to some HW.

A probabilistic black box model does not readily allow the generation of corner cases, since the model will, by definition, generate signals from the (learned) distribution. If the bug-revealing corner case is in the distribution support, it will take a long time to produce.[3] And if it is not in the distribution support, it can never be produced. While HW randomized testing allows us to vary the distribution from which the input transactions are generated, a learned black box model does not provide as much freedom. That's because the distribution is learned from data and changing it might reduce the validity of the generated signals. Therefore, if, say, functional coverage of the device is too low, we need another model altogether to generate the directed tests that finally get us above the coverage threshold required for sign-off.

Using a probabilistic model also complicates model checking, since the model checker must now return probabilistic answers. Probabilistic model checkers do exist for certain types of models like Discrete-Time Markov Chains [8], and it remains to be seen whether they scale to the input models we are interested in.

VII. Conclusion

The testing and validation of life-critical medical devices requires the conducting of a clinical trial. These trials are lengthy and costly, and the idea of using computer models of the physiology to help plan and execute the trial is very attractive. Physiological modeling happens at a very high level, and has unique challenges that distinguish it from other modeling efforts aimed at device debugging. In particular, the need for physiological validity, which has no counterpart in testing of consumer electronics, requires learning model parameters from real patient data. While these models can be re-used for device testing, several challenges complicate their direct usage to achieve the usual sing-off criteria of coverage.

References

[1] Ask The ICD. http://asktheicd.com, 2015. Accessed on 10/11/2015.

[2] M. R. Gold et al. Prospective comparison of discrimination algorithms to prevent inappropriate ICD therapy: Primary results of the Rhythm ID Going Head to Head Trial . *Heart Rhythm*, 9(3):370 – 377, 2012.

[3] Some techniques like importance sampling might help.

[3] M.E. Josephson. *Clinical Cardiac Electrophysiology*. Lippincot Williams and Wilkins, 2008.

[4] J. P. Daubert et al. Inappropriate Implantable Cardioverter-Defibrillator Shocks in MADIT II: Frequency, Mechanisms, Predictors, and Survival Impact . *Journal of the American College of Cardiology*, 51(14):1357 – 1365, 2008.

[5] A. J. Moss et al. Reduction in Inappropriate Therapy and Mortality through ICD Programming. *New England Journal of Medicine*, 367(24):2275–2283, 2012.

[6] S. Saria, A. Duchi, and D. Koller. Learning deformable motifs in continuous time series data. *International Joint Conference on Artificial Intelligence*, 2011.

[7] Lawrence Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.

[8] M. Kwiatkowska, G. Norman, and D. Parker. PRISM 4.0: Verification of probabilistic real-time systems. In G. Gopalakrishnan and S. Qadeer, editors, *Proc. 23rd International Conference on Computer Aided Verification (CAV'11)*, volume 6806 of *LNCS*, pages 585–591. Springer, 2011.